Aiding Pronoun Translation with Co-reference Resolution

Aiding Pronoun Translation with Co-Reference Resolution

Ronan Le Nagard and Philipp Koehn
University of Edinburgh
Edinburgh, United Kingdom
s0678231@sms.ed.ac.uk, pkoehn@inf.ed.ac.uk

Abstract

We propose a method to improve the translation of pronouns by resolving their co-reference to prior mentions. We report results using two different co-reference resolution methods and point to remaining challenges.

1 Introduction

While machine translation research has made great progress over the last years, including the increasing exploitation of linguistic annotation, the problems are mainly framed as the translation of isolated sentences. This restriction of the task ignores several discourse-level problems, such as the translation of pronouns.

Pronouns typically refer to earlier mention of entities, and the nature of these entities may matter for translation. A glaring case is the translation of the English *it* and *they* into languages with grammatical gender (as for instance, most European languages). If *it* refers to an object that has a male grammatical gender in the target language, then its translation is a male pronoun (e.g., *il* in French), while referring to a female object requires a female pronoun (e.g., *elle* in French).

Figure 1 illustrates the problem. Given a pair of sentence such as

The window is open. It is blue.

the translation of *it* cannot be determined given only the sentence it occurs in. It is essential that we connect it to the entity *the window* in the previous sentence.

Making such a connection between references to the same entity is called co-reference resolution, or anaphora resolution.[1] While this problem has motivated significant research in the field of natural language processing, the integration of co-reference resolution methods into machine translation has been lacking. The recent wave of work on statistical machine translation has essentially not moved beyond sentence-level and has not touched co-reference resolution.

Our approach to aiding pronoun translation with co-reference resolution can be outlined as follows. On both training and test data, we identify the anaphoric noun of each occurrence of *it* and *they* on the source side (English). We then identify the noun's translation into the target language (in our experiments, French), and identify the target noun's grammatical gender. Based on that gender, we replace *it* with *it-masculine, it-feminine* or *it-neutral* (ditto for *they*). We train a statistical machine translation system with a thusly annotated corpus and apply it to the annotated test sentences.

Our experiments show some degree of success of the method, but also highlight that current co-reference resolution methods (we implemented Hobbs and Lappin/Laess) have not yet achieved sufficient performance to significantly reduce the number of errors in pronoun translation.

2 Related Work

2.1 Co-Reference and Machine Translation

The problem of anaphora resolution applied to machine translation has not been treated much in the literature. Although some papers refer to the problem, their content is mostly concerned with the problem of anaphora resolution and speak very little about the integration of such an algorithm in the bigger theme of machine translation.

Mitkov et al. [1995] deplore the lack of study of the question and try to address it with the implementation of an anaphora resolution model and its integration into the CAT2 translation system [Sharp, 1988], a transfer system that uses an ab-

[1] In the context of pronouns, anaphora resolution and co-reference resolution are identical, but they differ in other contexts.

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 252–261,
Uppsala, Sweden, 15-16 July 2010.

The window is open. **It** *is blue.*	*La fenêtre est ouverte.* **Elle** *est bleue.*	CORRECT
The window is open. **It** *is black.*	*La fenêtre est ouverte.* **Il** *est noir.*	WRONG
The oven is open. **It** *is new.*	*Le four est ouverte.* **Elle** *est neuve.*	WRONG
The door is open. **It** *is new.*	*La porte est ouverte.* **Elle** *est neuve.*	CORRECT

Figure 1: Translation errors due to lack of co-reference resolution (created with Google Translate).

stract intermediate representation. The anaphora resolution step adds additional features to the intermediate representation.

Leass and Schwall [1991] present a list of rules to be implemented directly into the machine translation system. These rules seem to work mostly like a dictionary and are checked in a priority order. They state what should be the translation of a pronoun in each special case. Being specific to the problem of translating anaphors into Korean, these are of little interest to our current work.

2.2 Co-Reference : Syntactic Method

The first work on the resolution of pronouns was done in the 1970s, largely based on a syntactic approach. This work was based on empirical data and observations about natural languages. For example, Winograd [1972] uses the notion of co-reference chains when stating that if a single pronoun is used several times in a sentence or a group of adjunct sentences, all instances of this pronoun should refer to the same entity.

Others have also stated that antecedents of a pronoun should be found in one of the n sentences preceding the pronouns, where n should be small [Klapholz and Lockman, 1975]. Hobbs [1978] showed that this number was close to one, although no actual limit could be really imposed.

In work by both Hobbs [1978] and Winograd [1972], the resolution of pronouns also involves a syntactic study of the parse tree of sentences. The order with which candidate antecedents are prioritized is similar in both studies. They first look for the antecedent to be a subject, then the direct object of a noun and finally an indirect object. Only thereafter previous sentences are checked for an antecedent, in no particular order, although the left to right order seems to be preferred in the literature as it implicitly preserves the order just mentioned. Winograd uses focus values of noun phrases in sentences to choose the appropriate antecedent.

Hobbs also refers to the work by Charniak [1972] and Wilks [1975] for the problem of anaphora resolution. However, they do not offer a complete solution to the problem. For this reason Hobbs [1978] is often considered to be the most comprehensive early syntactic study of the problem, and as such, often used as a baseline to evaluate anaphora resolution methods. We use his work and comment on it in a later section.

Another approach to anaphora resolution is based on the centering theory first proposed by Grosz et al. [1995]. Brennan et al. [1987] propose an algorithm for pronoun resolution based on centering theory. Once again, the entities are ranked according to their grammatical role, where subject is more salient than existential constructs, which are more salient than direct and indirect objects. Walker [1998] further improves the theory of centering theory for anaphora resolution, proposing the idea of cache model to replace the stack model described originally.

Another syntactic approach to the problem of co-reference resolution is the use of weighted features by Lappin and Leass [1994] which we present in more details in a further section. This algorithm is based on two modules, a syntactic filter followed by a system of salience weighting. The algorithm gathers all potential noun phrase antecedents of a pronoun from the current and close previous sentences. The syntactic filter then filters out the ones that are unlikely to be antecedents, according to different rules, including general agreement rules. The remaining candidate noun phrases are weighted according to salience factors. The authors demonstrate a higher success rate with their algorithm (86%) than with their implementation of the Hobbs algorithm (82%).

2.3 Co-Reference : Statistical Approach

Machine Learning has also been applied to the problem of anaphora resolution. Ng [2005] gives a survey of the research carried out in this area.

The work by Aone and Bennett [1995] is among the first in this field. It applies machine learning to anaphora resolution on Japanese text. The authors use a set of 66 features, related to both the referent itself and to the relation between the referent and

its antecedent. They include "lexical (e.g. category), syntactic (e.g. grammatical role), semantic (e.g. semantic class), and positional (e.g. distance between anaphor and antecedent)" information.

Ge et al. [1998] also present a statistical algorithm based on the study of statistical data in a large corpus and the application of a naive Bayes model. The authors report an accuracy rate of 82.9%, or 84.2% with the addition of statistical data on gender categorization of words.

In more recent work, Kehler et al. [2004] show a move towards the use of common-sense knowledge to help the resolution of anaphors. They use referring probabilities taken from a large annotated corpus as a knowledge base.

2.4 Shared Tasks and Evaluation

Although a fairly large amount of research has been done in the field, it is often reported [Mitkov et al., 1995] that there does not yet exist a method to resolve pronouns which is entirely satisfactory and effective. Different kinds of texts (novel, newspaper,...) pose problems [Hobbs, 1978] and the field is also victim of lack of standardization.

Algorithms are evaluated on different texts and large annotated corpora with co-reference information is lacking to check results. A response to these problems came with the creation of shared tasks, such as the MUC [Grishman and Sundheim, 1996] which included a co-reference subtask [Chinchor and Hirschmann, 1997] and led to the creation of the MUC-6 and MUC-7 corpora.

There are other annotation efforts worth mentioning, such as the ARRAU corpus [Poesio and Artstein, 2008] which include texts from various sources and deals with previous problems in annotation such as anaphora ambiguity and annotation of information on agreement, grammatical function and reference. The Anaphoric Bank and the Phrase Detectives are both part of the Anawiki project [Poesio et al., 2008] and also promise the creation of a standardized corpus. The first one allows for the sharing of annotated corpora. The second is a collaborative effort to annotate large corpora through the Web. In its first year of use, the system saw the resolution of 700,000 pronouns.

3 Method

The method has two main aspects: the application of co-reference to annotate pronouns and the subsequent integration into statistical machine translation. We begin our description with the latter aspect.

3.1 Integration into Machine Translation

English pronouns such as *it* (and *they*) do not have a unique French translation, but rather several words are potential translations. Note that for simplicity we comment here on the pronoun *it*, but the same conclusions can be drawn from the study of the plural pronoun *they*.

In most cases, the translation ambiguity cannot be resolved in the context of a single sentence because the pronoun refers to an antecedent in a previous sentence. Statistical machine translation focuses on single sentences and therefore cannot deal with antecedents in previous sentences. Our approach does not fundamentally change the statistical machine translation approach, but treats the necessary pronoun classification as a external task.

Hence, the pronoun *it* is annotated, resulting in the three different surface forms presented to the translation system: *it-neutral*, *it-feminine*, *it-masculine*. These therefore encode the gender information of the pronoun and each of them will be match to its corresponding French translation in the translation table.

An interesting point to note is the fact that these pronouns only encode gender information about the pronouns and omit number and person information. This has two reasons.

Firstly, study of the lexical translation table for the baseline system shows that the probability of having the singular pronoun *it* translated into the plural pronouns *ils* and *elles* is 10 times smaller than the one for the singular/singular translation pair. This means that the number of times a singular pronoun in English translates into a plural pronoun in French is negligible.

The other reason to omit the cases when a singular pronoun is translated into a plural pronoun is due to the performance of our algorithm. Indeed, the detection of number information in the algorithm is not good enough and returns many false results which would reduce the performance of the final system. Also, adding the number agreement to the pronoun would mean a high segmentation between all the different possibilities, which we assumed would result in worse performance of the translation system.

Once we have created a way to tag the pronouns with gender information, the system needs to learn

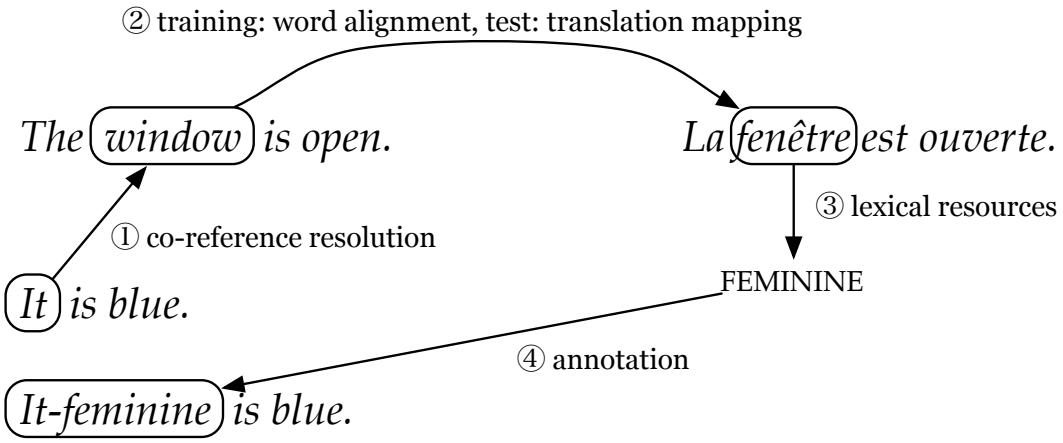

Figure 2: Overview of the process to annotate pronouns: The word *it* is connected to the antecedent *window* which was translated as *fenêtre*, a feminine noun. Thus, the pronoun is annotated as *it-feminine*.

the new probabilities that link the source language pronoun to the target language pronouns. That is all instances of *it* in the training data, which can be found at any position in the corpus sentences, should be replaced by one of its three declension. However, it is important to stress that the gender information that should be encoded in the English corpus is the one which corresponds to the gender of the French translation of the antecedent.

In order to find the correct gender information for the pronoun, we execute the co-reference resolution algorithm on the English text which returns the antecedent of the pronoun (more on this in the next section). Note that we are not interested in the English gender of the antecedent, but in gender of its translation.

Thus, we need to detect the French translation of the English antecedent. For the training data, we rely on the word alignment that is produced as a by-product of the training of a statistical machine translation system. For the test data, we rely on the implicit word mapping performed during the translation process.

Note that this requires in practice the translation of all preceding sentences before we can annotate the current sentence. To avoid this practical burden in our experiments, we simply use the mapping in the baseline translation. The performance of the sentence alignment (88

Once the French word is obtained, it is used as the input of a module which returns the gender of the entity in French. This is then used to replace the original pronoun with the new gendered pronoun.

The entire process is illustrated in Figure 2.

3.2 The Hobbs Algorithm

The Hobbs algorithm is considered to be the baseline algorithm for co-reference resolution. The algorithm uses the syntactic parse tree of the sentences as input.

The algorithm traverses the parse tree and selects appropriate candidate referents to the pronoun. It goes up sentence nodes and checks all NP nodes encountered for agreement with the pronoun. The order in which the algorithm traverses the tree ensures that some priorities are respected, to make sure the most probable antecedent is returned first. By doing this, the algorithm tends to enforces some of the constraints that apply to co-reference [Jurafsky et al., 2000]. The recency constraint is enforced thanks to the order in which the algorithm traverses the sentences and both the binding and grammatical role constraints are enforced by the use of the syntactic tree and Part-Of-Speech tags on the words.

Because the algorithm only uses the parse tree of the sentences, the semantic meaning of words is completely omitted in the process of selecting candidate antecedents and no knowledge is required except for the implicit knowledge contained within agreement features.

As mentioned earlier, the Hobbs algorithm goes up the tree from the given pronoun to the top of the tree and stops at each sentence or noun node on its way. In each of these nodes, it performs breadth first search of the sub tree and returns any noun phrase node encountered as a potential antecedent. If the antecedent is genuine (according to gender, number, and person agreement), it is returned.

In case no antecedent was found in the current sentence, the algorithm goes back up in the text, looking at each sentence separately, in a left-to-right breadth first fashion. This ensures that the subject/object/indirect object priorities and hierarchy are respected. Again, if a candidate NP has matching agreement features, it is returned as the antecedent of the pronoun. Otherwise the algorithm goes one sentence higher.

The original algorithm uses limited knowledge because it assumes that:

- Dates do not move.
- Places do not move.
- Large fixed objects don't move.

This add limited semantic restrictions for the antecedent chosen. Indeed, if the pronoun is followed by a motion verb, the antecedent could not be a date, a place or a large fixed object. However, as Hobbs states himself, those constraints help little since they do not apply in most cases.

3.3 The Lappin and Leass Algorithm

Lappin and Leass [1994] proposed an anaphora resolution algorithm for third person pronouns and lexical anaphors. It is based on slot grammar and uses syntax combined with a system of weights to select the appropriate antecedent of a pronoun. The implementation of the algorithm we deal with here is fairly different from the one presented in the original paper, and is largely inspired from the JavaRAP implementation [Qiu et al., 2004].

The first important variation was mentioned earlier and concerns the application of co-reference resolution to machine translation. We concentrate in this work on the resolution of third person pronouns, and we omit reflexive pronouns (*itself*, *themselves*) (referred to as lexical anaphora in some works).

Another variation comes from the use of the Collins parser [Collins, 2003]. Although work on the original algorithm uses McCord's Slot Grammar parser [McCord, 1990], work on JavaRAP shows that rules can be created to simulate the categories and predicates used in slot grammar. Also, Preiss [2002] evaluates the use of different parsers for the Lappin and Leass algorithm, showing that performance of the algorithm is not related to the performance of the parser itself. The JavaRAP implementation uses a Charniak parser, which performs worse than the Collins parser in Preiss' research.

For these reasons and in order to allow for reuse of the code used previously in the implementation of the Hobbs algorithm, the input to the Lappin and Leass algorithm is text parsed with the Collins parser.

It should be noted that the Lappin and Leass algorithm (also called RAP for Resolution of Anaphora Procedure) has been used in the original research for the application of machine translation.

The algorithm processes sentence by sentence, keeping in memory the information regarding the last four sentences. In the first step of the algorithm, all noun phrases (NPs) are extracted and classified. Definite and indefinite NPs are separated, and pleonastic pronouns are segregated from other pronouns.

The notion of salience is very important in RAP, as it allows the algorithm to choose between competing NPs. All candidate NPs are given a "salience weighting", which represents the importance and visibility of the phrase in the sentence, and in relation to the pronoun that is being resolved.

Salience weighting is based on the syntactic form of the sentence and the value for an NP is calculated through the contribution, or not, of different salience factors, to which weights are associated. This calculation ensures that different importance will be given to a subject noun phrase in a sentence, and a noun phrase that is embedded in another or that represents the indirect object of a verb.

There are a number of salience factors such as sentence recency, subject emphasis, existential emphasis, accusative emphasis, etc. Each factor is associated with a predefined weight.

Once the weight of each candidate has been calculated, the algorithm uses syntactic information to filter out the noun phrases that the pronoun is unlikely to refer to. This includes agreement and other checks.

The list of candidate NPs obtained after this processing is then cleared of all NPs that fall under a given threshold. The original algorithm then deals with singular and plural pronouns in different ways. The JavaRAP implementation however does not use these differences and we refer the reader to Lappin and Leass' paper for further information.

Finally, the candidate NPs mentioned in the previous list are ranked according to their salience

weights and the highest scoring one is returned as the antecedent of the pronoun. In case several NPs have the same salience weight, the one closest to the pronoun is returned.

3.4 Pleonastic It

English makes an extensive use of the pronoun *it* in a pleonastic fashion. That is, many times, *it* is considered to be structural and does not refer to any entity previously mentioned. The following are examples of pleonastic uses of *it*:

- *It is raining.*
- *It seems important that I see him.*
- *The session is opened, it was announced.*

Being able to discriminate the use of a structural *it* from the use of a referential use of *it* is very important for the success of the co-reference algorithm. Indeed, resolving a pleonastic *it* will be a waste of time for the algorithm, and more importantly, it will increase the chance of errors and will result in poorer performances. Moreover, the pleonastic *it* is most times translated masculine in French, meaning any other resolution by the algorithm will yield errors.

In the past, the importance given to the detection of the pleonastic use of *it* has varied from author to author. As an example, Rush et al. [1971], in their work on automatic summarization, only mentioned the problem. Others formed a set of rules to detect them, such as Liddy et al. [1987] with 142 rules, or Lappin and Leass [1994] who propose a very restricted set of rules for the detection of the structural *it*.

Paice and Husk [1987] carried out extensive research on the topic and their paper defines various categories for the pronoun *it* as well as proposing a set of rules that allow to differentiate when the pronoun *it* is used as a relational pronoun or as a pleonastic pronoun.

Their method categorise words according to the presence of given words around the pronoun it. They distinguish constructs such as *it VERB STATUS to TASK* ; construct expressing doubt containing words such as *whether*, *if*, *how* ; parenthetical it such as *it seems*, *it was said*. The original article identifies seven categories for pleonastic pronouns.

Since their own results showed a success rate of 92.2% on a test section of the LOBC corpus and the implementation of their technique yields results similar to the implementation of a machine learning technique, this method seemed appropriate for our purpose.

4 Experiments

In this section, we comment on the tools used for the implementation of the algorithms, as well as support tools and corpora.

The implementation of both of the algorithms was done using the Python programming language, which was chosen for its simplicity in processing text files and because it is the language in which the Natural Language Toolkit is developed.

The Natural Language Toolkit (NLTK) is a suite of Python modules used for research into natural language processing. We mostly used its Tree and ParentedTree modules which enable the representation of parse trees into tree structures. NLTK also includes a naive Bayes classifier, which we used in association with the names corpus in order to classify proper names into gender categories according to a set of features. We also use NLTK for its named entity capacities, in order to find animacity information of entities.

English sentences were annotated with the MXPOST Part of Speech tagger and the Collins syntactic parser.

The Lefff lexicon, introduced by Sagot et al. [2006] was used to get agreement features of French words. It contains over 80,000 French words,[2] along with gender and number information.

We used the open source Moses toolkit [Koehn et al., 2007] and trained standard phrase-based translation models.

As training data, we used the Europarl corpus [Koehn, 2005], a commonly used parallel corpus in statistical machine translation research. While there are also commonly used Europarl test sets, these do not contain sentences in sequence for complete documents. Instead, we used as test set the proceedings from October 5, 2000 - a set of 1742 sentences from the held-out portion of the corpus. We translated the test set both with a baseline system and a system trained on the annotated training data and tested on an annotated test set.

[2]The original version version of the lexicon is available from http://www.labri.fr/perso/clement/lefff/.

	Word	Count
English singular	*he*	17,181
	she	4,575
	it	214,047
French singular	*il*	187,921
	elle	45,682
English plural	*they*	54,776
French plural	*ils*	32,350
	elles	16,238

Table 1: Number of sentences in the training corpus containing third person personal pronouns.

Truth	Method	
	Pleonastic	Referential
Pleonastic	**42**	20
Referential	19	**98**

Table 2: Detection of pleonastic pronouns

5 Results

5.1 Corpus Statistics for Pronouns

Personal pronouns are among the most frequent words in text. In the training corpus of 1,393,452 sentences, about a 6th contain third person personal pronouns. See Table 1 for detailed statistics.

The English pronoun *it* is much more frequent than *he* or *she*. For both languages, the masculine forms are more frequent than the feminine forms.

There are then a total of 233,603 sentences containing a third person pronoun in French, and 235,803 sentences containing a third person pronoun in English. This means that over 2,000 of those pronouns in English do not have equivalent in French. Similarly for plural: A total of 48,588 sentences contain a plural pronoun in French, against 54,776 in English. That shows that over 6,000 of the English ones are not translated into French.

5.2 Detection of the Pleonastic *it*

We checked, how well our method for pleonastic *it* detection works on a section of the test set. We achieved both recall and precision of 83% for the categorization of the referential *it*. For details, please see Table 2.

5.3 Translation Probabilities

Let us now examine the translation probabilities for the annotated and un-annotated pronouns. Details are given in Table 3.

correct annotation	33/59	56%
correct translation:		
annotated	40/59	68%
correctly annotated	27/33	82%
baseline	41/59	69%

Table 4: **Translation Results:** On a manually examined portion of the test set, only 33 of 59 pronouns are labeled correctly. The translation results of our method does not differ significantly from the baseline. Most of the correctly annotated pronouns are translated correctly.

In the baseline system, both *it* and *they* have a strong translation preference for the masculine over the feminine form of the French pronoun. *It* translates with probability 0.307 to *il* and with probability *0.090* to *elle*. The rest of the probability mass is taken up by the NULL token, punctuation, and a long tail of unlikely choices.

For both the Hobbs and the Lappin/Laess algorithm, the probability distribution is shifted to the desired French pronoun. The shift is strongest for the masculine marked *they*, which prefers the masculine *ils* with 0.431 over the feminine *elles* with 0.053 (numbers for Hobbs, Lappin/Laess numbers are 0.435 and 0.054, respectively).

Feminine marked pronouns now slightly prefer feminine French forms, overcoming the original bias. The neutrally marked pronouns shift slightly in favor of masculine translations.

The pronoun *they-neutral* appears in 12,424 sentences in the corpus, which all represent failed resolution of the co-reference. Indeed, French does not have neutral gender and the plural third person pronoun is never pleonastic. These results therefore show that a lot of noise is added to the system.

5.4 Translation Results

The BLEU scores for our method is almost identical to the baseline performance. This is not surprising, since we only expect to change the translation of a small number of words (however, important words for understanding the meaning of the text).

A better evaluation metric is the number of correctly translated pronouns. This requires manual inspection of the translation results. Results are given in Table 4.

While the shift of the translation probabilities

Unannotated

English	French	p
it	*il*	0.307
it	*elle*	0.090
they	*ils*	0.341
they	*elles*	0.130

Hobbs

English	French	p
it-neutral	*il*	0.369
it-neutral	*elle*	0.065
it-masculine	*il*	0.230
it-masculine	*elle*	0.060
it-feminine	*il*	0.144
it-feminine	*elle*	0.168
they-neutral	*ils*	0.344
they-neutral	*elles*	0.102
they-masc.	*ils*	0.435
they-masc.	*elles*	0.053
they-feminine	*ils*	0.208
they-feminine	*elles*	0.259

Lappin and Laess

English	French	p
it-neutral	*il*	0.372
it-neutral	*elle*	0.064
it-masculine	*il*	0.211
it-masculine	*elle*	0.051
it-feminine	*il*	0.142
it-feminine	*elle*	0.156
they-neutral	*ils*	0.354
they-neutral	*elles*	0.090
they-masc.	*ils*	0.431
they-masc.	*elles*	0.054
they-feminine	*ils*	0.207
they-feminine	*elles*	0.255

Table 3: **Translation probabilities.** The probabilities of gender-marked pronouns are shifted to the corresponding gender in the two cases the text was annotated with the co-reference resolution methods mentionned earlier.

suggests that we are moving the translation of pronouns in the right direction, this is not reflected by the sample of pronoun translations we inspected. In fact, the performance for our method is almost identical to the baseline (68% and 69%, respectively).

One cause for this is the poor performance of the co-reference resolution method, which labels only 56% of pronouns correctly. On this sub-sample of correctly annotated pronouns, we achieve 82% correct translations. However, the baseline method also performs well on this subset.

6 Conclusion

We presented a method to aid pronoun translation for statistical machine translation by using co-reference resolution. This is to our knowledge the first such work.

While our method works in principle, the results are not yet convincing. The main problem is the low performance of the co-reference resolution algorithm we used. The method works well when the co-reference resolution algorithm provides correct results.

Future work should concentrate on better co-reference algorithms. The context of machine translation also provides an interesting testbed for such algorithms, since it offers standard test sets for many language pairs.

7 Acknowledgements

This work was supported by the EuroMatrixPlus project funded by the European Commission (7th Framework Programme).

References

C. Aone and S.W. Bennett. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 122–129. Association for Computational Linguistics Morristown, NJ, USA, 1995.

S. E. Brennan, M. W. Friedman, and C. J. Pollard. A centering approach to pronouns. In *Proceedings of the 25th annual meeting on Association for Computational Linguistics*, pages 155–162, 1987.

E. Charniak. *Toward a model of children's story comprehension*. MIT, 1972.

N. Chinchor and L. Hirschmann. MUC-7 coreference task definition, version 3.0. In *Proceedings of MUC*, volume 7, 1997.

M. Collins. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637, 2003.

N. Ge, J. Hale, and E. Charniak. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–170, 1998.

R. Grishman and B. Sundheim. Message understanding conference-6: A brief history. In *Proceedings of the 16th conference on Computational Linguistics-Volume 1*, pages 466–471. Association for Computational Linguistics Morristown, NJ, USA, 1996.

B. J. Grosz, S. Weinstein, and A. K. Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.

J. R. Hobbs. Resolving Pronoun References. *Lingua*, 44:339–352, 1978.

D. Jurafsky, J. H. Martin, A. Kehler, K. Vander Linden, and N. Ward. *Speech and language processing*. Prentice Hall New York, 2000.

A. Kehler, D. Appelt, L. Taylor, and A. Simma. The (non) utility of predicate-argument frequencies for pronoun interpretation. In *Proc. of HLT-NAACL*, volume 4, pages 289–296, 2004.

D. Klapholz and A. Lockman. Contextual reference resolution. *American Journal of Computational Linguistics, microfiche 36*, 1975.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, Phuket, Thailand, September 2005.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

S. Lappin and H.J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):561, 1994.

Herbert Leass and Ulrike Schwall. An Anaphora Resolution Procedure for Machine Translation. Technical Report Report 172, IBM Germany Science Center, Institute for Knowledge Based Systems, 1991.

E. Liddy, S. Bonzi, J. Katzer, and E. Oddy. A study of discourse anaphora in scientific abstracts. *Journal of the American Society for Information Science*, 38(4):255–261, 1987.

Michael C. McCord. Slot grammar: A system for simpler construction of practical natural language grammars. In *Proceedings of the International Symposium on Natural Language and Logic*, pages 118–145, London, UK, 1990. Springer-Verlag. ISBN 3-540-53082-7.

R. Mitkov, S. K. Choi, and R. Sharp. Anaphora resolution in Machine Translation. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, TMI'95*, 1995.

V. Ng. Machine learning for coreference resolution: From local classification to global ranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, page 164. Association for Computational Linguistics, 2005.

C. D. Paice and G. D. Husk. Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun. *Computer Speech & Language*, 2(2):109–132, 1987.

M. Poesio and R. Artstein. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2008.

M. Poesio, U. Kruschwitz, and J. Chamberlain. ANAWIKI: Creating anaphorically annotated resources through Web cooperation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, volume 8. Citeseer, 2008.

Judita Preiss. Choosing a parser for anaphora resolution. In *4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, pages 175–180. Edi cões Colibri, 2002.

Long Qiu, Min yen Kan, and Tat seng Chua. A public reference implementation of the rap anaphora resolution algorithm. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 291–294, 2004.

J. E. Rush, R. Salvador, and A. Zamora. Automatic abstracting and indexing. II. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. *Journal of the American Society for Information Science and Technology*, 22(4):260–274, 1971.

B. Sagot, L. Clément, E. V. de La Clergerie, and P. Boullier. The Lefff 2 syntactic lexicon for French: architecture, acquisition, use. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2006.

Randall Sharp. CAT2 – implementing a formalism for multi-lingual MT. In *2nd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*, pages 3–6, 1988.

M. A. Walker. Centering, anaphora resolution, and discourse structure. *Centering theory in discourse*, pages 401–435, 1998.

Y. Wilks. A preferential, pattern-seeking, semantics for natural language inference. *Words and Intelligence I*, pages 83–102, 1975.

T. Winograd. Understanding natural language. *Cognitive Psychology*, 3(1):1–191, 1972.